Encyclopedia of Library and Information Sciences, Third Edition
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t917508581

Internet Archive

Marilyn Rackley[a]
[a] Library, Harvard University, Cambridge, Massachusetts, U.S.A.

Online publication date: 09 December 2009

To cite this Article Rackley, Marilyn(2010) 'Internet Archive', Encyclopedia of Library and Information Sciences, Third Edition, 1: 1, 2966 — 2976

Internet Archive

Marilyn Rackley
Library, Harvard University, Cambridge, Massachusetts, U.S.A.

Abstract
This entry describes the history of the Internet Archive from its founding in 1996 to its two billion page crawl in 2007. It describes the key individuals and organizations involved in the Archive's work and the technological innovations that make the Archive possible, such as the ARC file format, Heritrix, and the Wayback Machine. The focus of this entry is primarily the Internet Archive's Web archiving activities and collections, but it also briefly discusses the Archive's other activities and the impact it has had on the fields of library and information science and the public in general.

INTRODUCTION

The Internet Archive (http://www.archive.org) has captured the imaginations of information professionals and the Internet-using public alike. As a pioneer in the field of Web archiving, it offers the largest and broadest collection of historic Web sites in the world and can lay claim to a substantial collection of other digital cultural heritage objects as well. Although often best known for its Wayback Machine, a portal to its Web site archives, the Archive and its cofounder and director, Brewster Kahle, are active proponents of making all human knowledge available to everyone in digital formats and of preserving that knowledge for future generations; and the creation of the Wayback Machine is only one aspect of their pursuit of this goal.

Indeed, the Internet Archive has played both a direct and an indirect role in preserving human knowledge. It has played a direct role through its collecting activity and technology development and it has played an indirect role through its inspiration of similar collecting and preservation efforts. Moreover, the Archive collaborates with dozens of other institutions engaged in similar pursuits and promotes policies and practices that increase the free, public availability of cultural heritage materials.

This entry describes the history of the organization, including its mission and policies; the technology it uses to achieve its mission; its collections and services; and the impact it has had on the areas of Web archiving and digital preservation in general.

HISTORY

Mission

The nonprofit Archive was founded in 1996 as a response to a rapidly growing problem—the disappearance of content from the World Wide Web. Estimates of the average life span of a Web site vary, but the Internet Archive's Frequently Asked Questions page gives 77 days as the figure.[1] As the Internet became increasingly prominent in all areas of human activity, it became correspondingly important to preserve it. Archive cofounder Brewster Kahle likened the potential disappearance of digital content from the Internet to other great catastrophic historical and cultural losses; "[m]anuscripts from the library of Alexandria in ancient Egypt disappeared in a fire. The early printed books decayed into unrecognizable shreds. Many of the oldest cinematic films were recycled for their silver content. Unfortunately, history may repeat itself in the evolution of the Internet—and its World Wide Web."[2]

To combat this problem, Kahle and Bruce Gilliat, two technologists, established the Internet Archive in San Francisco, California. They made access to and preservation of Web content the organization's primary mission, and their vision for the organization was ambitious, aiming at nothing short of an archive of the entire World Wide Web. Because no other individual or organization had previously attempted to archive Web content on such a large scale, the Archive needed not only to create a vast collection of this ephemeral content, but also to invent the very means by which that content could be collected and subsequently preserved.

In June of 2007, the state of California recognized the Internet Archive as a library, making it eligible for Federal funding and symbolically allying it with institutions that have a long tradition of preserving and providing access to knowledge. The Internet Archive encourages and promotes this association by, for example calling user accounts "library cards." The Archive shares the ideals and aspirations of libraries and other similar organizations. Early in its history, the Library of Congress tapped the organization to be its partner in creating collections based on important events in U.S. history, such as the

2000 national elections and the terrorist attacks of September 11, 2001, conferring prestige on the Archive and signaling a shared agenda between the collaborators.

However, the Archive also sets itself apart from traditional libraries through its unique collections. Since its creation, the Internet Archive has made "universal access to all human knowledge" the foundation of all its activities. Because it is neither a collection of material selected to serve the needs of a specific community nor a collection preserving the historical memory of a specific organization, it has been said that "the Internet Archive might be described as a true archive, seeking to collect and preserve the entire Web, past, present, and future."[3] In the years since its foundation, it has grown from an organization that archives Web content to an organization that promotes the free exchange of information in a variety of forms and formats. It does so by making content available online in digital formats and by developing tools and services that help others to make content in their own collections more widely available. Fig. 1 shows the home page of the Internet Archive in late 2007. The content highlighted on the home page gives an indication of the variety of materials in the Archive's collections.

Founders and Significant Contributors

The individual most commonly associated with the Internet Archive is Brewster Kahle (b. 1960). Kahle was one of the founders and still leads the organization as its director. He began his career in the technology industry; after receiving a degree in Computer Science and Engineering from Massachusetts Institute of Technology (MIT) in 1982, Kahle's first position was at Thinking Machines Corporation, a supercomputer manufacturer in Massachusetts. At Thinking Machines, he participated in the development of Wide Area Information Servers (WAIS). Kahle left Thinking Machines to start WAIS, Inc. in 1992, which AOL then purchased in 1995.

Since selling WAIS, Kahle has brought his background in technology and business to bear on the issue of the role of technology in society. Kahle characterizes the Internet as a democratic communication medium and he directs much of his time toward defending and promoting what he views as an egalitarian medium. In 2005, the American Academy of Arts and Sciences elected Kahle as a fellow. The Academy and its members promote service, scholarship, and public engagement in a wide range

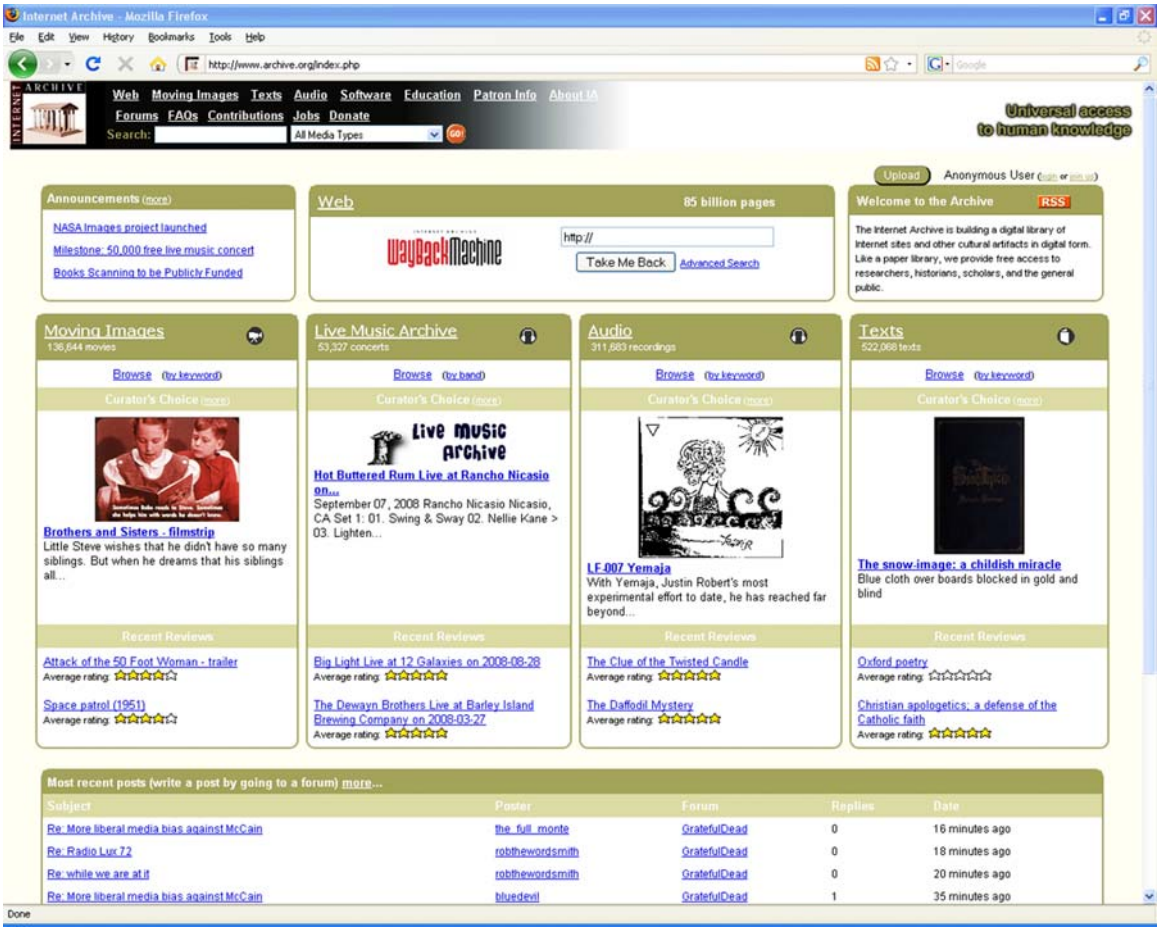

Fig. 1 The Internet Archive home page (http://www.archive.org) on September 23, 2008.

of disciplines; and Kahle is an active public participant in contemporary discussions of technology's impact on education, the economy, culture, and the law. In one demonstration of his commitment to open access to information, he has supported challenges to the extension of copyright protection and to the copyright status of orphan works, acting both through the Internet Archive and as an individual. (The Archive filed an amicus brief in the case of Eldred vs. Ashcroft, which challenged the constitutionality of the Copyright Term Extension Act of 1998. Kahle, with Richard Prelinger, also brought suit against the United States to challenge the constitutionality of U.S. copyright law. Both cases were heard by the Supreme Court, and in both cases, the plaintiffs were unsuccessful in loosening copyright restrictions.)

Bruce Gilliat founded the Internet Archive and Alexa Internet with Brewster Kahle. Gilliat has remained at Alexa and is currently the Chief Executive Officer. Although Gilliat is not directly involved with the Archive's activities, through Alexa Internet, he continues to work on improved access to Web content and helps preserve it via donations of content for the Archive's collections.

Another individual who has made important contributions to the Archive is the current president of the board of directors, Rick Prelinger. Prelinger is now best known for his archive of over 60,000 ephemeral films, a collection which he began in 1983, but he is also a filmmaker and writer. The collection of educational, promotional, amateur, and industrial films bearing his name is the largest of its kind. In 2002, the Library of Congress acquired the collection, but Prelinger continues to collect in the area. Like Kahle, Prelinger is an enthusiastic advocate for the preservation of and public access to cultural heritage materials.

A key institutional partner in the Internet Archive's Web archiving activity is Alexa Internet, a for-profit organization that has provided the content for the Archive's Web collection since the Archive's inception. Alexa Internet was founded in late 1996 to catalog the Web. The company collects, aggregates, and analyses data on Web content and Web usage. It gathers information on Web content through its Web crawls, which run continuously, and it gathers information on Web usage from the use of the Alexa Toolbar. The company has sold this information to other companies, including other Internet application creators, such as Netscape. Alexa Internet has been the defendant in more than one lawsuit alleging violation of privacy laws in its data collection and storage practices. Although Amazon purchased the company in 1999, it continues to donate copies of its Web crawls to the Internet Archive.

TECHNOLOGY

The work of the Internet Archive requires a specific set of tools for the capture, storage, and provision of access to its collections. Because many of these tools did not exist in 1996 when the organization first began its work, it has been responsible, in many cases, for designing and creating these tools itself, either on its own or with partners.

Capture Technology

The Internet Archive currently uses the open source Heritrix Web crawler to find and capture Web sites. The Archive began developing Heritrix in 2003, using the Java programming language. Prior to 2003, the Archive relied solely on Alexa Internet's proprietary crawler to collect Web files. Heritrix's development has involved the Archive, the International Internet Preservation Consortium, and other partner libraries and institutions.

The basic requirements for the original crawler used by Alexa Internet and the Internet Archive have remained unchanged even while specific crawlers have progressed. First, the crawler must be as polite as possible, obeying any instructions on crawl behavior found in a site's robots.txt file and avoiding aggressive crawl behavior that might disrupt other site traffic. Second, the crawler must be able to function on more than one machine at a time. Thirdly, the files captured from the Web must be aggregated into larger files for ease of management and access.[4]

Beginning with specific Universal Resource Locators (URLs) known as seeds, the crawler finds documents (files) available over the Internet and downloads them to the Archive's servers. The crawler parses the files looking for references (links) to other files represented by URLs and then adds those URLs to the list of files to retrieve. If these reference paths are relative, the paths are made absolute before being added to the list. As the crawler retrieves each file in the list, it repeats this process of searching for references to additional files and verifies that the additional files have not already been captured. This process is continuously operating to acquire multiple snapshots of the Web as it grows and changes.

However, capturing the Web in snapshots, which take place at varying intervals, cannot result in a complete picture of the Web over time. Files and content added to a Web page and taken off in between Alexa crawls of that page will never be present in the archive. Because of the method the crawler uses to find new documents to retrieve, files to which other files frequently link will be captured more often than files to which other files rarely link. The number and timing of the Internet Archive snapshots available for any given file collected from the Internet will vary and will often present only an incomplete picture of any changes affecting that document over time.

As Heritrix and other crawlers have improved over the years, their ability to capture larger and more complicated Web sites has improved, as they become better able to parse the references in each file. Future releases of Heritrix will continue to improve crawl results and will include further duplication reduction, increased control for

curators, and advanced prioritization features.[5] Nonetheless, with the technology currently available, much of the Internet remains hidden from Web crawling robots, and thus outside of the Archive's collecting scope. Databases cannot be captured by the Archive using Heritrix, nor can it collect files that are password protected. In addition, Heritrix has difficulty parsing links found in many of the more complicated Web programming technologies, such as JavaScript and Flash, which are becoming increasingly popular with Web designers. A user of the Internet Archive's Web collection may find that only the home page of a site is available in the archive because the site's navigation tools were built in JavaScript, which the crawler was not able to parse, and therefore the files making up the rest of the site were not added to the crawler's list of documents to retrieve.

Storage and Preservation

In order to efficiently store the billions of files that make up the Web pages in its collection, the Archive created the Archive File Format (ARC), which uses the .arc extension. The requirements for this file format were that the captured files would be self-identifying, that it could store files retrieved via any type of network protocol; and that it would allow the aggregation of multiple archive files into larger files.[6] If the captured files were self-identifying, there would be no need for separate index files to assist in the retrieval of each individual file. If the file format could handle files retrieved through different protocols, it could handle any files available through the Internet, and not only those using the hypertext transfer protocol used, for instance, on the World Wide Web. Finally, if the system was not required to handle billions of small files, it would be able to use its storage and retrieval resources more efficiently.

Satisfying these requirements, each ARC file contains multiple files retrieved during a crawl, each with a header containing elements of metadata about the file and its retrieval. These elements include the file name (its URL), its size, its content type, the date and time of retrieval, and the name of the organization that retrieved it. The Archive stores the products of its Web crawls in ARC files holding approximately 100 MB of data each. Because of the ease with which archived files can be found within an ARC file, the files needed to view an archived Web page or site can be spread among multiple ARC files.

For the first 3 years of its existence, the Archive used tape to store archived files. However, the tapes were unable to efficiently handle a large volume of access requests and the need for a scalable storage solution became apparent as the number of access requests increased. Consequently, the Internet Archive designed the Petabox to store and process large amounts of data. As its name indicates, the Petabox can hold one petabyte of data, which is the equivalent of a million gigabytes. To house its massive archive, the organization needed inexpensive and efficient storage that was easy to maintain and would support the Archive's back-up and mirroring system. After developing the storage, the Internet Archive has let a third-party reproduce it for use in many other organizations.

One of the challenges to the long-term viability of any archive is the preservation of its collections. The Internet Archive recognizes the need to maintain copies of its collections in locations that are geographically separated from each other. Mirror sites exist in Alexandria, Egypt, and in Amsterdam, the Netherlands.

Access—The Wayback Machine

The Wayback Machine (http://www.archive.org/web/web.php) is now almost synonymous with the Internet Archive, but it was not until October 2001 that this simple browser-based user interface was introduced as the primary way to access the Archive's Web collections.

The Wayback Machine interface allows users to find a specific instance of a Web page. A user enters a URL into the Wayback Machine search field and is "taken back" to a results page listing the dates on which that page was captured (see Fig. 2). The user then selects the version she wishes to view by clicking on any available date.

Fig. 3 shows one of the earliest iterations of the Archive's home page as seen in the Wayback Machine before any collections were publicly available.

Fig. 4 shows how the home page has changed after a few years of the organization's existence, with a sizeable collection of archived Web sites freely available.

Additionally, the Wayback Machine allows users to browse the Internet through space and time from any given starting page. One of the most significant features of the application is its ability to rewrite hyperlinks to refer back to archived files as opposed to "live" files; links are rewritten when a user requests a page from the archives and not when the files are archived. With this capability, the user is able to link from the archived version of one page to the archived version of another page through the page's original links, replicating the interconnectedness of the "live" Web in the archive environment.

An example from the Internet Archive's home page will serve to illustrate how the Wayback Machine works. The URL for an August 15, 2000 version of the Internet Archive's home page in the Wayback Machine is http://web.archive.org/web/20000815054438/http://www.archive.org/. In addition to an indication that the user is in the domain of the Internet Archive's Web archive collection (http://web.archive.org), this URL includes the original URL for the resource being displayed (http://www.archive.org), as well as the date and time of its capture (August 15, 2000, 5:44:38). If a user clicks on a link from this instance of the home page to another page in the same site, such as the "About" link, she will be taken to a page with a URL of http://web.archive.org/web/20000815072836/www.archive.org/about/index.html, that is to an archived "About" page. The URL allows the user to know that they

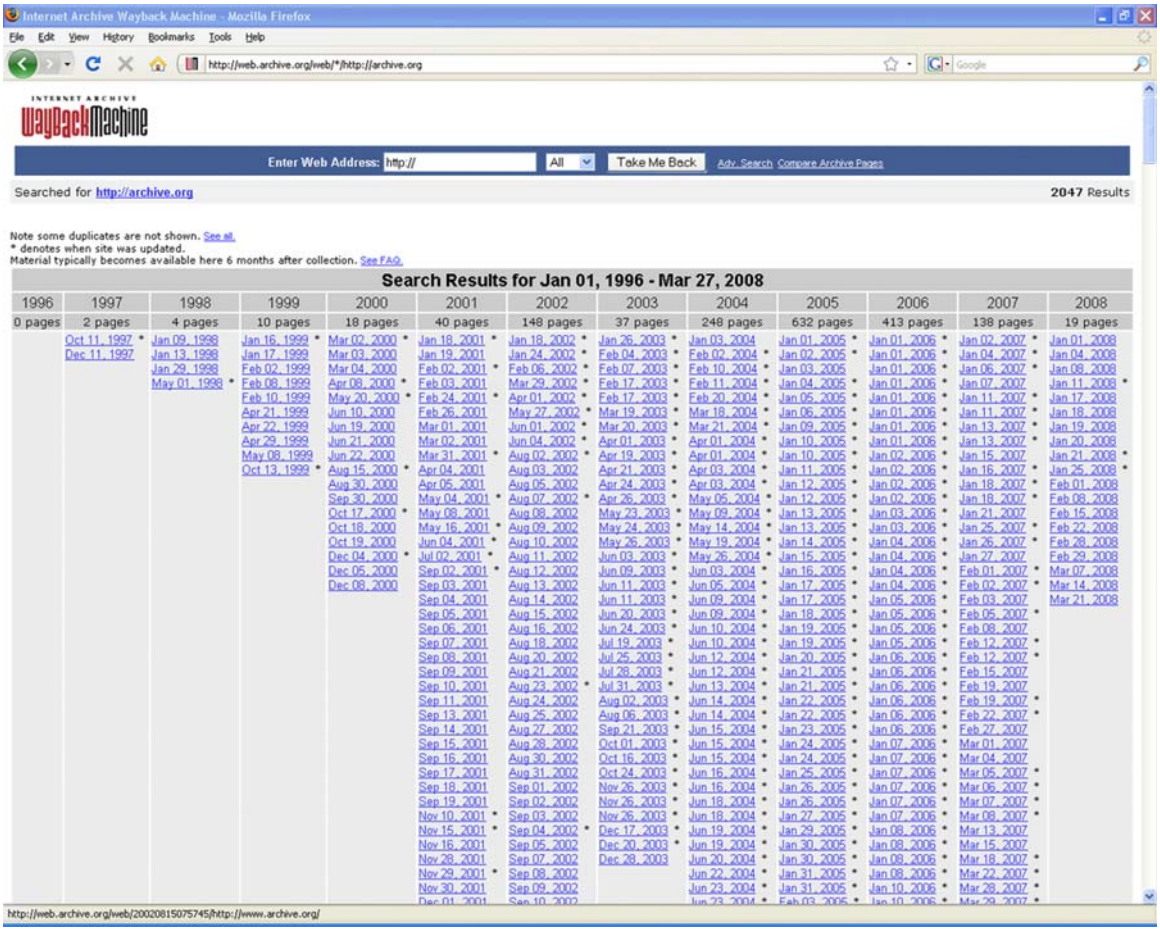

Fig. 2 The Wayback Machine results page for the URL http://archive.org (The Internet Archive home page) on September 23, 2008 (http://web.archive.org/web/*/http://archive.org).

are not on the "live" Web because she is still in the web.archive.org domain, that she is looking at a page that was originally found at http://www.archive.org/about/index.html, and that the page was captured on August 15, 2000 at 7:28:36. When the Archive's home page was displayed in the Wayback Machine, the "About" link was rewritten to point to the instance of the original link that is closest in capture time to the instance of the home page being displayed.

Similarly, if the user chooses to follow one of the links on the "About" page to a page on another site, such as the link to the UNESCO page about the new Library of Alexandria, she would be taken to a page in the archive represented by the URL http://web.archive.org/web/20000818203230/www.unesco.org/webworld/alexandria_new/. The page displayed then would represent a different space and time than the page from which the user first began navigating in the Wayback Machine, because it was originally from a different domain and was captured on August 18, 2000 at 20:32:30, approximately 3 days after the Archive's home page. Thus the user is able to navigate through the archived Web just as on the "live" Web, traveling from one domain to another, but she is also able to navigate through time by viewing files captured at different times.

The Internet Archive presents snapshots of files captured from the Internet as the Alexa crawlers follow available hyperlinks. This capture method prevents the Archive from presenting a site, domain, or other group of files as they existed at one time. Of the examples given above, none of the pages were captured at exactly the same time. Consequently, in this example, the user would be viewing only an approximation of what the Internet Archive site and the sites it linked to looked like on a specific date in August 2000 at a specific time.

While the primary means of retrieving archived Web sites through the Wayback Machine is by URL, the Archive is developing more advanced search options for the Web site archives, in particular a full-text search engine. Currently, there is an advanced search interface for the Web archive allowing users to limit their URL-based search by date, and a structured search interface is available for other portions of the Archive's collections, making possible searches by such fields as, date, creator, collection, and media type.

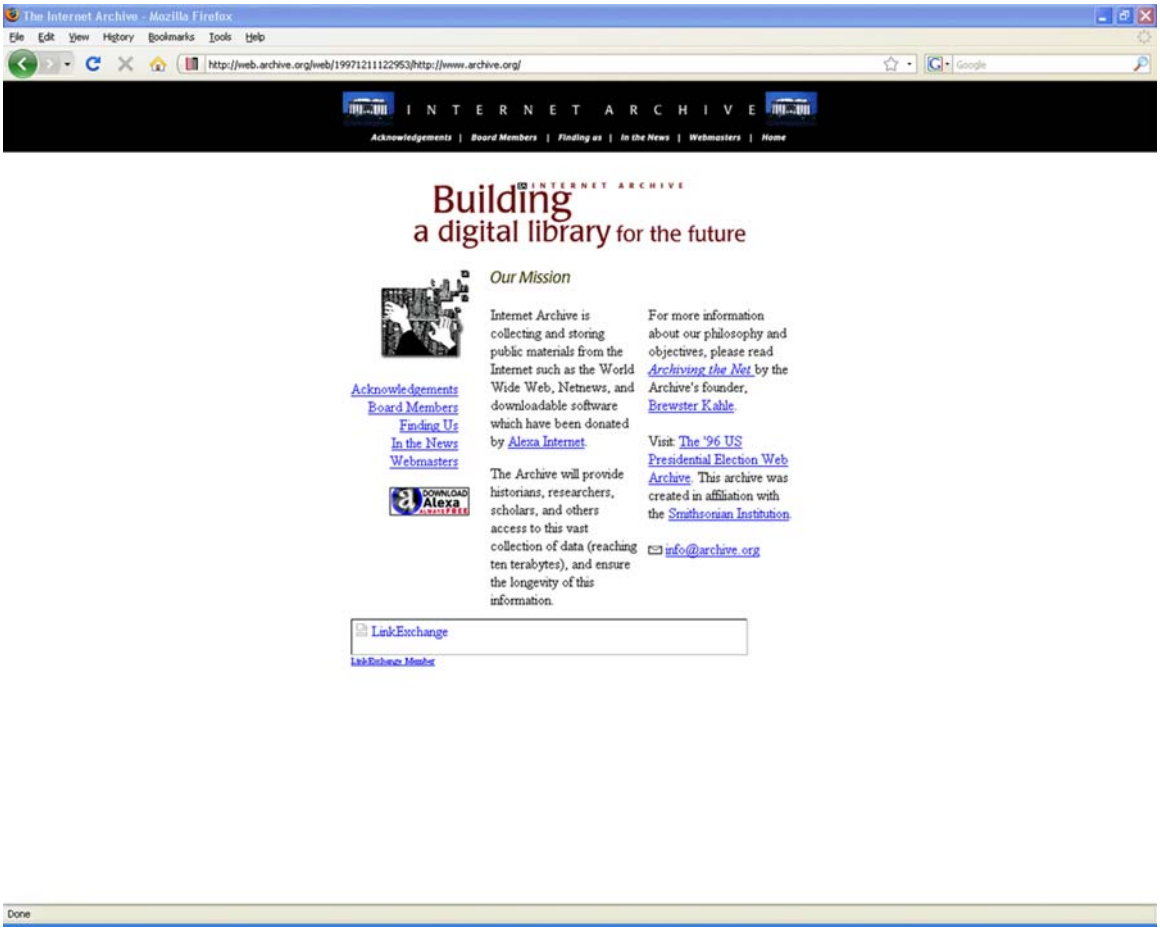

Fig. 3 The Internet Archive home page on December 11, 1997 as seen in the Wayback Machine (http://web.archive.org/web/19971211122953/http://www.archive.org/).

Like Heritrix, the Wayback Machine is an open source application, created by the Internet Archive, but also used by other institutions engaged in Web archiving activity. One significant drawback to the Wayback Machine is the lack of clues indicating to the user that she is viewing an archived Web page. The only way to identify the instance being presented is through the URL, and users must pay close attention to the URLs as they browse through the collection in order not to lose their way (or even in some cases, to find themselves suddenly on the "live" Web). However, browsing the historic Web through the Wayback Machine is only one way of accessing the Archive's vast collection. Users interested in applying data mining techniques to analyze older Web pages are permitted to create their own applications for viewing and exploring the collection.

COLLECTION AND SERVICES

Web Pages

The Archive's first and most significant collection is its collection of archived Web pages. The Alexa Internet crawler continually crawls the Internet, capturing new files and new versions of older files. When the Archive accessions these files, it is collecting documentation of the changes in individual Web sites and pages over time and changes in the Web as a whole.

This collection has generally been characterized by its breadth, rather than its depth, especially in its earliest accessions. The characteristics of the Web collection are the product of the Archive's philosophy regarding the nature of the Web and the manner in which the material was collected. According to Kahle, the Web as a whole is "a self-documenting, self-cataloging machine"; the content of the Web includes within it its own description and the means of its discovery.[7] Thus, the interconnectedness of all Web content necessitates an archiving strategy that aims to collect and preserve as much content as possible, without regard to the intrinsic merit of any given site. This view of the Web does not consider each file, page, or site as a separate entity to be specifically captured in its entirety, so, for example, the Archive would not point its crawler to a specific site, such as http://www.archive.org, and attempt to gather all the files found in that domain and not the files found on other domains.

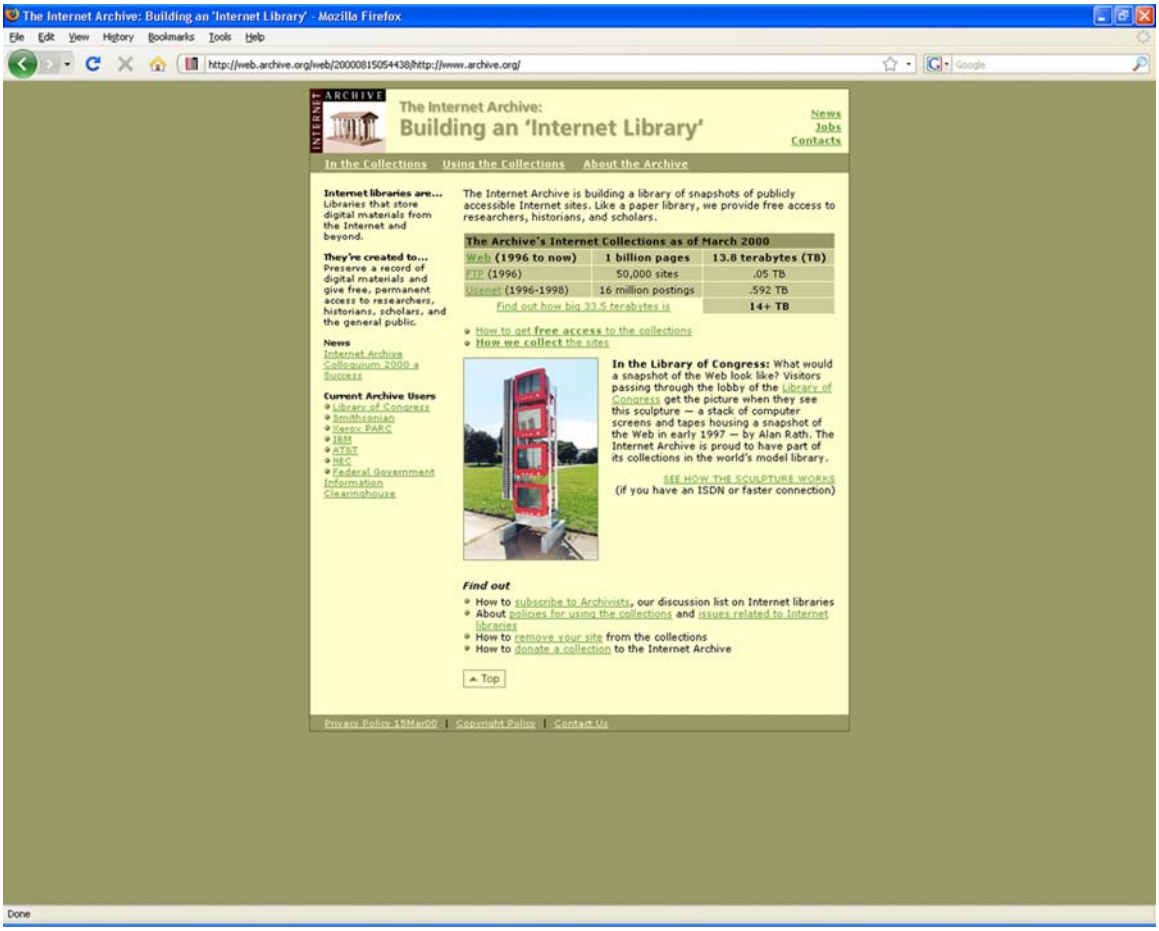

Fig. 4 The Internet Archive home page on August 15, 2000 as seen in the Wayback Machine (http://web.archive.org/web/20000815054438/http://www.archive.org/).

Instead, the crawler would be instructed to start with that URL as a seed and to follow all links found within the files on that domain even if they pointed to other domains. The crawler instructions might also specify a limited number of files to capture from any one domain or simply limit the amount of time spent retrieving documents. In this way, the crawler would collect files from as many domains as possible without ever necessarily collecting any domain or site in its entirety. The result would be a representation of the many sites existing on the Web at a given point in time, but would not be an exact replica of every file available through the Internet at that time.

Another factor limiting the content of the archives is the simple fact that the crawler must be aware that a site, page, or file exists to be able to retrieve. Alexa Internet's source of information about the existence of these documents comes from individuals using the Alexa Toolbar and from the links within documents already captured. If the Alexa crawler has not been made aware of the existence of certain files, it does not crawl them and they will not be added to the Archive's collection. Indeed, research has suggested that the Alexa crawler's link discovery method has resulted in Web pages hosted in certain countries to be considerably underrepresented in the Archive's collections.[8]

The Archive's Web collection has changed over the last decade as the technology needed to create it has improved, but the essence of its collecting policy remains the same. The organization aims to preserve as broad a snapshot of the Internet as possible. It does, though, recognize that it potentially collects materials, especially Web content, whose owners and copyright holders may not want them included in the Archive. Regarding such content, the Archive's policy is to remove the files from the collection at the author's or publisher's request. The Archive's crawlers also respect any instructions (found in robots.txt files) from Web site owners not to crawl their pages, and blocks future access to any such pages captured before the instructions were added. The Archive relies on its "polite" crawler and its "opt-out" policy to protect copyright owners and does not actively seek permission to collect and preserve the objects in its collections.

The organization does not appraise the materials it collects and does not undertake to censor any materials that viewers may consider objectionable.

Although by the end of 2006, the Archive had already collected over 85 billion pages, in June of 2007, it began a two billion page crawl to create a comprehensive, global snapshot of the Web at that time. This project was supported by a grant from the Mellon Foundation. Prior to the crawl's commencement, cultural institutions from around the world were able to submit their URLs for inclusion. The "Around the World in Two Billion Pages" collection was made publicly available at the end of 2007.

Despite the Archive's ambitious collecting goals, their achievement has been limited by the capture technologies and the associated methodologies the organization uses to collect Web-based documents, as well as its own collecting policies. Many documents simply cannot be identified and crawled by the Alexa spider. The Web collection does not include Internet-based databases, e-mail, discussion forum postings, and other types of content. Other documents that are identified and can be crawled are archived only when and because other documents link to them and not when their content is changed. Some documents will never be captured because of the file owner's instructions to the Alexa crawler not to capture the file, while still other documents might be retroactively removed from the archive to respect the wishes of copyright owners. Consequently, the Archive can neither truly reproduce an accurate and complete version of the Internet as it existed at any given point in time, nor can it provide even one version of all the documents available on the Internet since the organization's collecting began.

Furthermore, while capture and replay technology continues to improve at a rapid pace, certain types of dynamic Web page scripting, especially client-side scripting, such as JavaScript, are difficult and sometimes impossible to properly capture and replay. Because of these constraints, archived Web pages and even entire archived sites may not always accurately reflect the "look and feel" of the originals.

Other Digital and Digitized Content

Since 1999, the collecting scope of the Archive has included other forms of digital content, in addition to Web-based digital content. The Archive now boasts of over 300,000 digitized texts, 200,000 audio files, 100,000 video files, and 30,000 software-related materials. Like the Web archives, the materials in the other collections cover a wide variety of topics. While some materials are educational, others are closer to pure entertainment. Indeed, one of the most popular topics, if the Archive's users forum is any indication, is the band The Grateful Dead.

Much of this material, though hosted at the Archive, has actually come into the collections through donations from individuals and other institutions, including, for instance, the long-standing Project Gutenberg, which offers plain text versions of public domain books and some audio books, and the MIT's OpenCourseWare project, which has contributed lecture videos to the Archive. The Internet Archive offers to store and provide access to these files to further its goal of promoting and, to the extent possible, providing universal access to all knowledge.

Any user may add files to the collections through an Internet-based interface. Unlike traditional archives, the Internet Archive does not play an active role in choosing accessions to its collections, nor does it make any claim to guarantee the authenticity and reliability of these accessions. It also does not attempt to control access to or use of the materials.

Other Services

In 2005, the Internet Archive launched a service "designed for institutions that have been mandated to preserve content from the public Web but do not have the IT infrastructure or technical staff necessary to meet that mandate at the current time."[9] This service, named Archive-It, currently has over 50 partners, including state governments, universities, and nonprofit organizations. For an annual fee, partners are able to use Archive-It's Web-based application to create and manage their collections. The Internet Archive maintains the files making up the collections and provides access to them through a collection-specific version of the Wayback Machine.

The Internet Archive is also actively engaged in promoting access to and use of its content, especially in under-served communities. One notable example is its bookmobiles. Launched initially in the United States in 2002, the Internet bookmobile traveled across the country giving away small paperback children's books printed on demand, using the Archive's collection of digitized public-domain classics. The bookmobile was the foundation for a separate project, Anywhere Books, which created bookmobiles in Egypt, India, and Uganda. However, the Internet bookmobile is not currently traveling.

IMPACT

The pioneering efforts of the Internet Archive have been hugely influential in the digital preservation and access community, and its director Brewster Kahle is well-known inside and outside this community for his enthusiastic promotion of the Archive's goals. The organization's influence has been especially significant in the area of Web archiving, inspiring and providing models for other projects and programs around the world, but it has also stimulated debate about the preservation of digital objects and its collections have already been used in legal discovery and academic research.

Other Web and Digital Archives

The success of the Internet Archive has been the inspiration for many other large-scale Web archiving programs,

and it has collaborated with other nonprofit organizations and national institutions on some of those programs. In addition to providing a model for Web archiving activity, the Archive has provided a majority of the technology currently being used by other institutions to create Web archiving applications, programs, and collections around in the world.

In the United States, for instance, the open-source crawler Heritrix forms the basis of the crawler used by the Online Computer Library Center's (OCLC) Web Archiving Workbench application, and the Archive's collaboration with the Library of Congress was crucial in making possible the Library's MINERVA Web Archiving Project. In Europe, the National Archives of the United Kingdom uses the Archive to host its earliest Web site collections dating back to 1996. The Royal Library and the State and University Library of Denmark have also developed a Web archiving application, the NetarchiveSuite, around the Heritrix crawler.

Other efforts have benefited less directly from the work of the Internet Archive, but undoubtedly have been influenced by the Archive as the most prominent and oldest example of Web archiving and a successful example of the archiving of digital objects. For example, Scientists at the Los Alamos National Lab in New Mexico are using the ARC file format as the basis for their aDORe Archive storage of digital objects.

While building on the work of the Internet Archive, many cultural heritage institutions have recognized the limitations of the Archive's collections and have also recognized a need for their own active participation in the collection and preservation of important documents and records falling within their collecting scopes. Those organizations with a mandate to collect materials now found on the Internet, whether those materials are organizational records or informational documents, cannot rely on the Internet Archive to completely and accurately archive these materials. Instead they must work to develop collecting, preservation, and access solutions of their own or work directly with third-party service providers to ensure the survival of these documents. Thus, for instance, many state governments and universities in the United States have begun Web archiving programs in order to ensure that their own institutional Web-based records are preserved for the future, and many libraries are developing thematically related collections of materials archived from the Internet.

The International Internet Preservation Consortium

The International Internet Preservation Consortium (IIPC) was chartered in 2003, with the Internet Archive joining with 11 national libraries to address the challenge of preserving Internet content. The Internet Archive was the only nongovernmental charter member. The Consortium studies Web archiving practices and standards in the areas of harvesting, preservation, and access. The IIPC is also at work on a new file format for Web archiving, using ARC as a starting point. Called WARC, this new format will include a way to record actions taken to preserve the files.

The Open Content Alliance

The Open Content Alliance (OCA) was founded in 2005, with Yahoo. The goal of the OCA is to provide free access to global digital content. Although the Alliance includes commercial organizations, like Yahoo, Microsoft, and Adobe Systems, its purpose is to provide an alternative to commercial control of human knowledge. The majority of the members are university and other large research libraries, public, and private. Kahle envisions an organization that will empower libraries to work together to keep library content open to the public in response to his perception of a trend toward increased corporate control of that content.[10]

The primary focus of the Alliance's work is printed books and analog audio and video, no longer under copyright protection, rather than the born-digital materials that formed the foundation of the Internet Archive's collections. As part of the OCA, the Internet Archive offers to digitize library materials for only a small per page or per disc cost. The Alliance has not yet made any content available, but the plans for inaugural collections include materials from the University of California, the National Archives of the United Kingdom, O'Reilly Media, and the European Archive.

Copyright

When collecting materials found on the "live" Web, the Internet Archive does not gain explicit permission from their owners to copy and archive them. Instead it relies on its "opt-out" policy as described above and the doctrine of fair use, which has a well-established history in the world of information in analog formats, but becomes more contentious in the digital world, where it is often much easier to copy and widely disseminate copyrighted materials. The doctrine of fair use might allow the Internet Archive to make a copy of a Web site for noncommercial uses, but it is more questionable whether the doctrine still applies when the Archive makes that same copy available to millions of people via the Internet.

Thus far, the organization has not been involved in much significant litigation regarding copyright issues. Indeed the most significant lawsuit against the Archive to date involves not whether it has the right to collect Web pages but whether it failed to follow its own policy respecting crawl instructions by copyright owners.[11] Its policies and its status as a research and education resource have somewhat reduced the legal risk inherent in its mission. The Archive's actions are certainly an important test

of the fair use doctrine in the realm of digital preservation, but it is still far from clear what the rights of archivists, librarians, and other information professionals are with respect to collecting, preserving, and making accessible digital objects.

Court Cases

The Archive's Web collection has already been used to find evidence in both civil and criminal court cases. The Wayback Machine is becoming well-known among lawyers, particularly those who handle trademark and domain name litigation. Lawyers can use the Wayback Machine to find examples of trademark infringement even if the offender has already removed the trademarked material from the "live" site. Although using archived or cached Web pages as evidence is not unequivocally accepted in the legal community, this type of evidence has gained significant attention in the past several years for being of considerable assistance in the evidence discovery process. It is likely that there will be more questions about the reliability and authenticity of this information, especially when it is used in criminal cases.[12,13]

Scholarship

Researchers in the field of information and library science have already begun to use the Internet Archive's collections to perform historical research on the Web and its development. For example, in 2004, a paper on the accessibility of government Web sites for people with disabilities from 1997–2002 was presented at an Association for Computing Machinery (ACM) conference.[14] The authors used the Wayback Machine to view and test government Web sites from the past. Researchers at Cornell University are building the Cornell Web Library, which will use the collections of the Internet Archive but will provide researchers more technologically sophisticated ways of organizing and analyzing the collections without requiring them to be sophisticated programmers.[15]

CONCLUSION

The Internet Archive has been an early and prominent advocate for the preservation of the digital cultural heritage of civilization. Even if its own collections and collecting activity were insignificant, the role the organization, its founders, and other employees and collaborators have played in initiating and furthering public dialog about all aspects of access to information in the digital age are significant enough to earn the Internet Archive an important place in the history of cultural institutions. Its influence has touched large and small cultural heritage institutions in areas ranging from the very specific details of digital preservation (by designing a new preservation file format, for instance) to the very broad issue of access to information in digital forms.

Some critics have questioned the long-term viability of a cultural heritage institution founded by two technologists, but cultural heritage institutions founded by dedicated philanthropists are not without successful precedent in the analog world. Nonetheless, the organization will face many new challenges in the future if it is to preserve its collections. Some of these challenges will be technical in nature as the practical requirements of maintaining a large collection of digital files in a variety of formats change over time, but other challenges will require policy-related resolutions. Allegations of copyright violations have arisen, and given the nature of the Archive's "opt-out" collecting policy, there are likely to be more in the future.

Furthermore, the Archive may need to overcome limitations in its collecting methodologies, policies, and technologies if it is to build a truly comprehensive collection. The Internet Archive, like other organizations engaged in archiving the Web and other forms of digital content, faces four primary problems: the cultural problem (deciding what to save), the technical problem (what technology is necessary and viable), the economic problem (who will pay for it), and the legal problem (who owns the material and who can use it).[16] The Internet Archive has been finding bold solutions to these problems for over a decade in its quest to save the cultural heritage of civilization from catastrophic loss.

REFERENCES

1. Internet Archive Frequently Asked Questions, http://www.archive.org/about/faqs.php (accessed September 2007).
2. Kahle, B. Preserving the Internet. Sci. Am. **1997**, *276* (3), 82–83.
3. Kahle, B.; Lyman, P. Archiving digital cultural artifacts. D-Lib Mag. **1998**, *4* (7), http://www.dlib.org/dlib/july98/07lyman.html (accessed September 2007).
4. Burner, M. Crawling towards eternity: Building an archive of the World Wide Web. Web Tech. Mag. **1997**, *2* (5), 37–40, http://www.webtechniques.com/archives/1997/05/burner/ (accessed September 2007).
5. Mohr, G. IA/IIPC open source tools update. In 7th International Web Archiving Workshop, Vancouver, Canada, June 23, 2007, http://www.iwaw.net/07/mohr-iwaw07.pdf (accessed September 2007).
6. Burner, M.; Kahle, B. Arc file format, http://www.archive.org/web/researcher/ArcFileFormat.php (accessed September 2007).
7. Kahle, B. Editor's interview: The Internet Archive, an interview with Brewster Kahle. RLG DigiNews **2002**, *6* (3), http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070519:000006287741&reqid=3550#interview (accessed September 2007).
8. Thelwall, M.; Vaughan, L. A fair history of the Web? Examining country balance in the Internet. Archive. Libr. Inform. Sci. Res. **2004**, *26*, 162–176.

9. Archive-It Questions, http://www.archive-it.org/public/faq (accessed September 2007).
10. Albanese, A.R. Scan this book! Libr. J. **2007**, *132*, 32–35. August, http://www.libraryjournal.com/article/CA6466634.html (accessed September 2007).
11. Zeller, T. Keeper of expired web pages is sued because archive was used in another suit. New York Times. July 13, 2005, C9.
12. Kesmodel, D. Not fade away—Lawyers' delight: Old web material doesn't disappear; Wayback Machine and Google archive billions of pages, including deleted ones; playboy protests 'sex court.' Wall Street J. July 27, **2005,** A1.
13. Fagan, M. 'Can you do a wayback on that?' The legal community's use of cached web pages in and out of trial. Boston Univ. J. Sci. Technol. Law **2007**, *13* (1), 46–73.
14. Hackett, S.; Parmanto, B.; Zeng, X. Accessibility of Internet websites through time, In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility*, Atlanta, GA, October 18–20, 1994; 32–39.
15. Arms, W.; Aya, S.; Dmitriev, P.; Kot, B.; Mitchell, R.; Walle, L. A research library based on the historical collections of the Internet Archive. D-Lib Mag. **2006**, *12* (2), http://www.dlib.org/dlib/february06/arms/02arms.html#2 (accessed September 2007).
16. Lyman, P. Archiving the World Wide Web. In *Building a National Strategy for Preservation: Issues in Digital Media Archiving*; Council on Library and Information Resources and the Library of Congress: Washington, DC, 2002, http://www.clir.org/pubs/reports/pub106/web.html (accessed September 2007).

BIBLIOGRAPHY

1. Brown, A. *Archiving Websites: A Guide for Information Management Professionals*; London: Facet, 2006.
2. Kahle, B. Editor's interview: The Internet Archive, an interview with Brewster Kahle. RLG DigiNews **2002**, *6* (3), http://www.rlg.org/preserv/diginews/diginews6-3.html#interview.
3. Kahle, B. Preserving the Internet. Sci. Am. **1997**, *276* (3), 82–83.
4. Kahle, B.; Lyman, P. Archiving digital cultural artifacts. D-Lib Mag. **1998**, *4* (7), http://www.dlib.org/dlib/july98/07lyman.html.
5. Kimpton, M.; Ubois, J. Year-by-year: From an archive of the Internet to an archive on the Internet. In *Web Archiving*; Masanès, J., Ed.; Springer: Berlin, 2006; 201–212.
6. Masanès, J., Ed. *Web Archiving*; Springer: Berlin, 2006.
7. Mohr, G.; Stack, M.; Ranitovic, I.; Avery, D.; Kimpton, M. An Introduction to Heritrix, an Open Source Archival Quality Web Crawler. In Presentation at the 4th International Web Archiving Workshop, Bath, September 16, 2004, http://www.iwaw.net/04/Mohr.pdf.